APPROACH TO DATA CLUSTERING BASED ON MOLECULAR CHEMICAL REACTIONS WITH VARIOUS DISTANCE MEASURES

E.M. Markushin, G.Sh. Shkaberina, N.L. Rezova, L.A. Kazakovtsev

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121)

The problem of objects automatic classification

- There be a sample of research objects $A=\{A_1,...,A_N\}$, where N is the sample size.
- Each object is described using a set of M variables $Z_1,...,Z_M$. The set $Z=\{Z_1,...,Z_M\}$ can include variables of different types.
- **It is required** to form $k \geq 2$ classes (groups of objects).

  The number of classes can be preselected or determined automatically.

*Jain, A. K., & Dubes, R.C. (1988). *Algorithms for Clustering Data*. Prentice-Hall: New Jersey, USA, 96 – 101.

Known methods for solving problems of automatic classification of objects

- The k-means problem

$$\arg\min F(0.X_1, \dots, X_k) = \sum_{i=1}^{N} min_{j \in \overline{\{1,k\}}} \left\| X_j - A_j \right\|^2;$$

- The k-medoids model (Kaufman & Rousseeuw, 1987);
- The k-medians algorithm (Jain & Dubes, 1988);
- The CLARA (Clustering Large Applications) algorithm (Kaufman & Rousseeuw, 1990);
- The CLARANS (Clustering Large Applications based upon Randomized Search) algorithm (Ng & Han, 2002);
- The Variable Neighborhood Search (VNS) algorithms (Mladenovic & Hansen, 1997; Rozhnov et al., 2019);
- The agglomeration algorithms (Sun et al., 2014);
- o use genetic algorithms and other evolutionary approaches to improve local search results (Maulik & Bandyopadhyay, 2000; Krishna & Murty, 1999).

Known genetic algorithms for the k-means problem

- Complex algorithms: the genetic algorithms, neural networks (Holland, 1975), simulated annealing algorithm (Kirkpatrick et al., 1983).
- The genetic algorithm for solving the discrete p-median problem (Hosage & Goodchild, 1986);
- The genetic algorithm with solution coding in the form of a set of indices of network nodes (Bozkaya et al., 2002);
- The genetic algorithm with a special crossing procedure - a greedy (agglomerative) heuristic procedure (Alp et al., 2003);
- Many mutation methods can be used in genetic algorithms for k-means and similar problems (Kazakovtsev & Antamoshkin, 2014; Kwedlo & Iwanowicz, 2010);
- For the k-means and p-median problem, the mutation procedure, as a rule, changes one or more solutions, replacing some centers (Maulik & Bandyopadhyay, 2000; Krishna & Murty, 1999).

Clustering based on molecular chemical reactions

Atom 1		1	2	3	4	5	6	7	8		Point number
Atom 2		2	1	2	3	1	2	3	1		Cluster number

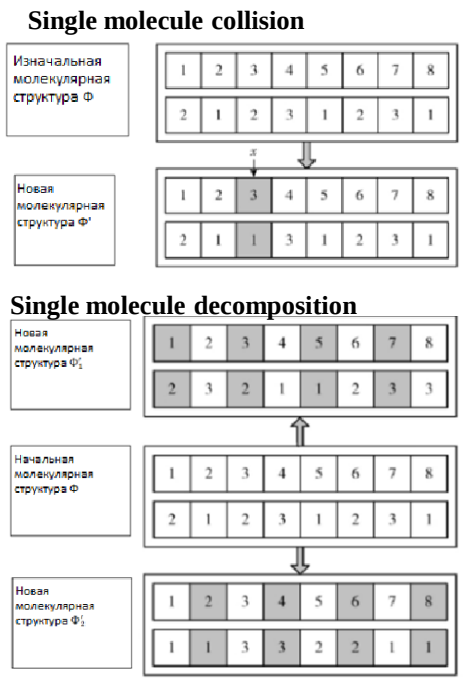
$$\operatorname{argmin} F(X_1,\ldots,X_k) = \sum \min_{j\in\{1,k\}} \|X_j - A_i\|^2$$

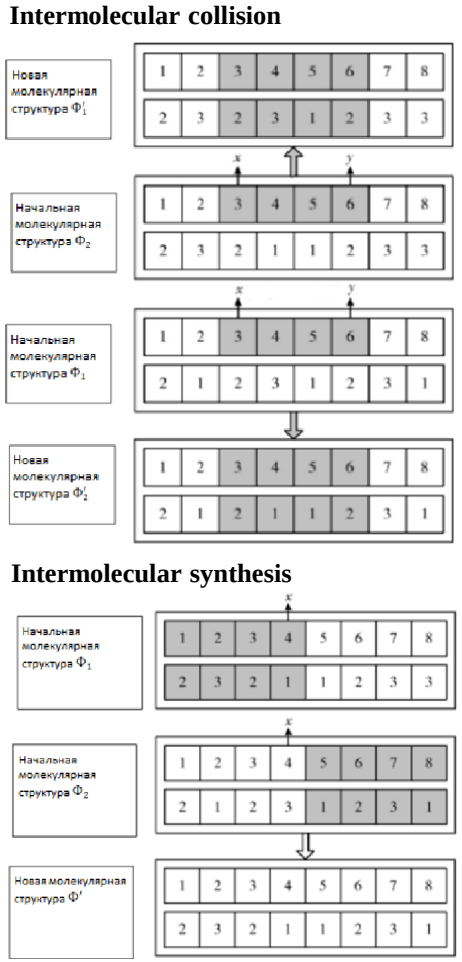
The distance measures:

The Euclidean distance (EuD): $$d(x,y) = \sqrt{\sum_{i=1}^{M} (x_i - y_i)^2}.$$

The squared Euclidean distance (SEuD): $$d(x,y) = \sum_{i=1}^{M} (x_i - y_i)^2.$$

The Manhattan distance (ManD): $$d(x,y) = \sum_{i=1}^{M} |x_i - y_i|.$$





Algorithm. KCR algorithm

Require: The initial data points $A_1,\ldots,A_N$, numbers of clusters k, the number of iterations of chemical reaction process n.

Step 1. Generation of a random initial solution $\Phi = \{X_1 \ldots X_k\}$;

Step 2. Apply the k-means algorithm to Φ to obtain a local optimum Φ';

Step 3. i=1;

Step 4. Apply the procedures of molecular chemical reactions for the individual Φ' to obtain a new solution Φ'':

Step 4.1. The generation of new clustering solution by Single molecule collision method from Φ'. Calculate the objective function (1) of a new molecule $F(\Phi_1'')$;

Step 4.2. The generation of new clustering solution by Single molecule decomposition method. Calculate the objective function (1) of a new molecule $F(\Phi_2'')$;

Step 4.3. Randomly the generation of new clustering solution $\Psi = \{X_1 \ldots X_k\}$. The generation of new clustering solution by Intermolecular collision method. Calculate the objective function (1) of a new molecule $F(\Phi_3'')$;

Step 4.4. Randomly the generation of new clustering solution $\Upsilon = \{X_1 \ldots X_k\}$. The generation of new clustering solution by Intermolecular synthesis method. Calculate the objective function (1) of a new molecule $F(\Phi_4'')$;

Step 5. Calculate $F(\Phi'')$, $F(\Phi'') = min(F(\Phi_1''), F(\Phi_2''), F(\Phi_3''), F(\Phi_4''))$;

Step 6. Apply the k-means algorithm to Φ'' to obtain a local optimum Φ''';

Step 7. i=i+1;

Step 8. IF $F(\Phi''')$>$F(\Phi')$ AND i<=n THEN $\Phi' \leftarrow \Phi''$; go to ШАГ 4 ELSE $\Phi' \leftarrow \Phi'''$

Step 9. Decode clustering solution Φ'.

Computational experiments. Synthetic (artificial) dataset

Parameter	k-means			KCR		
	EuD	SEuD	ManD	EuD	SEuD	ManD
min	100.2318	100.2318	102.5189	100.2318	100.2318	102.4449
max	100.2379	100.2379	160.385	100.2379	100.2369	159.8675
mean	100.2351	100.2353	115.9838		100.2334	115.8819
σ	0.003029	0.002945	24.29265	0.001683	0.001787	24.15255
V	0.003021	0.002938	20.94486	0.001679	0.001783	20.84239
R	0.006071	0.006071	57.86609	0.006071	0.00513	57.42261

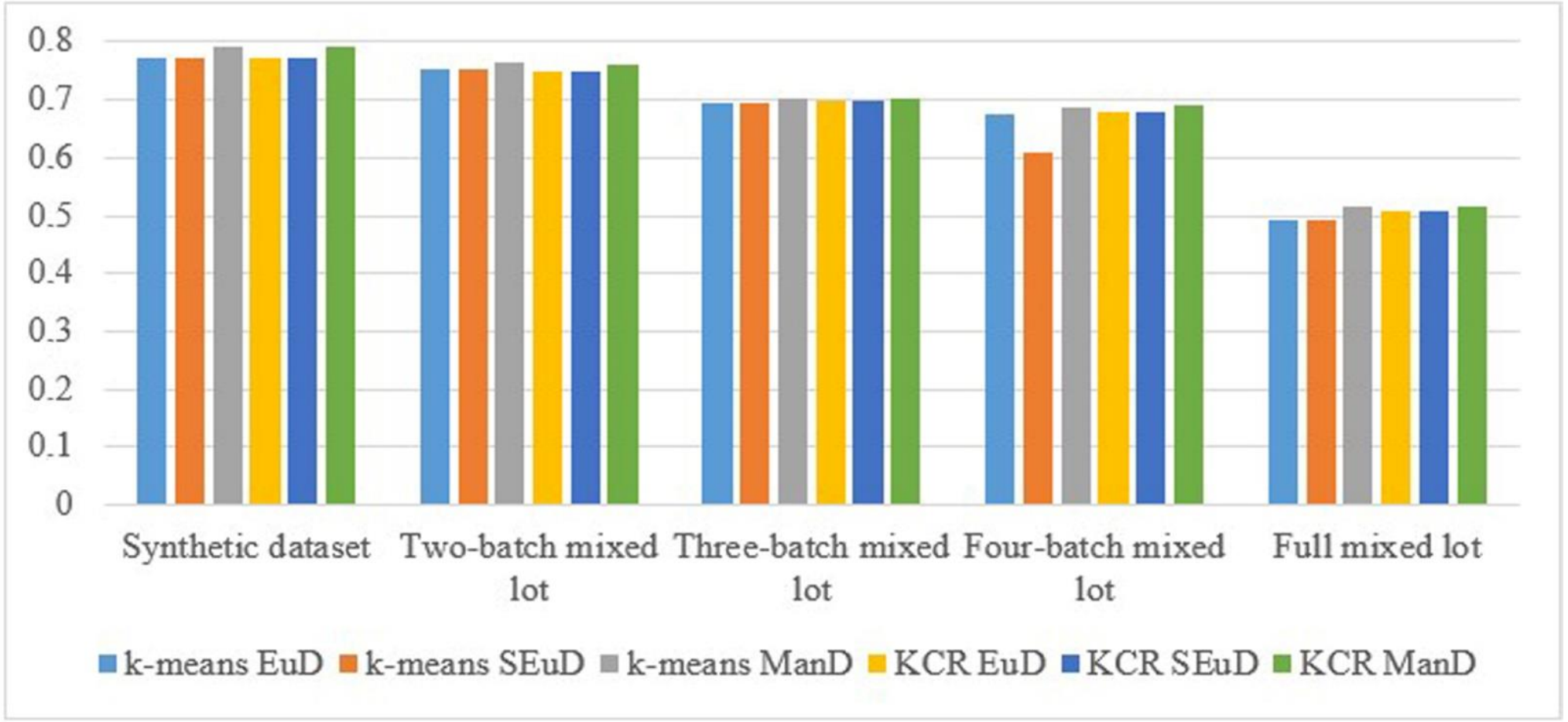
Computational experiments. Microchips 1526IE10_002

Parameter	k-means			KCR		
	EuD	SEuD	ManD	EuD	SEuD	ManD
Two-batch mixed lot (197 data points, 41 dimensions)						
min	0.026981824	0.026981824	0.027030361	0.026976786	0.026976786	0.026981824
max	0.026981824	0.026981824	0.027030361	0.026981824	0.026981824	0.027030361
mean	0.026981824	0.026981824	0.027030361	**0.026979305**	0.026980648	0.027027312
σ	0	0	0	2.52E-06	2.13E-06	9.20E-06
V	0	0	0	0.00933682	0.007897671	0.034030838
R	0	0	0	5.04E-06	5.04E-06	4.85E-05
Three-batch mixed lot (300 data points, 41 dimensions)						
min	0.048329262	0.048329262	0.048419577	0.048329262	0.048329262	0.048358367
max	0.048332067	0.048332067	0.04847852	0.048329262	0.048329262	0.048476331
mean	0.048329449	0.048329636	0.048465557	**0.048329262**	**0.048329262**	0.048460111
σ	7.00E-07	9.53E-07	1.98E-05	3.46E-18	3.46E-18	3.34E-05
V	0.001447527	0.001972639	0.040875312	7.16E-15	7.16E-15	0.068835581
R	2.80E-06	2.80E-06	5.89E-05	6.94E-18	6.94E-18	0.000117964

Computational experiments. Microchips 1526IE10_002

Parameter	k-means			KCR		
	EuD	SEuD	ManD	EuD	SEuD	ManD
Four-batch mixed lot (446 data points, 62 dimensions)						
min	0.011871273	0.011871273	0.012025967	0.011871273	0.011871273	0.012021142
max	0.016308472	0.016308472	0.016554985	0.011885287	0.011904385	0.013707829
mean	0.012022907	0.012906738	0.012343073	**0.011873588**	0.011874292	0.012097325
σ	0.000795859	0.001876659	0.001125694	4.39E-06	8.41E-06	0.000299143
V	6.619519898	14.54015291	9.12004666	0.0369922	0.070829331	2.472800111
R	0.0044372	0.0044372	0.004529018	1.40E-05	3.31E-05	0.001686687
Full mixed lot (3987 data points, 67 dimensions)						
min	0.118946137	0.118946137	0.119723176	0.118946137	0.118946575	0.1197271
max	0.164975905	0.157709299	0.159640787	0.143888254	0.157709241	0.14263887
mean	0.129590909	0.129384046	0.126030278	0.125914328	0.127245642	**0.124172035**
σ	0.012357251	0.011716148	0.010097741	0.008068167	0.010446673	0.007021172
V	9.53558456	9.055326403	8.01215488	6.407664014	8.209847419	5.654390752
R	0.046029769	0.038763163	0.039917611	0.024942118	0.038762666	0.02291177

Computational experiments. Accuracy of data clustering with various distance measures



Conclusions

- We proposed KCR algorithm for data clustering, which combines k-means algorithm and chemical reaction algorithms, could achieve an effective balance between local search capabilities and global exploration capabilities. The new algorithm improves the accuracy of solving the k-means problem.
- Computational experiments showed that in the vast majority of cases, minimal mean objective function value was demonstrated by KCR algorithm with Euclidean distance. However, using the KCR algorithm with Manhattan distance, in most cases, improves the accuracy of data clustering. In addition, the clustering accuracy increases with increasing number of points in the dataset.